The Fringes of Fitness: Statistical Profiling of CrossFit Games Open Winners

M. K., Toronto, BrainStation Data Science Boot Camp, 2021

Business Question

When exploring the fringes of fitness, we pursued a goal of finding profound ways to identify the best performers. It is common knowledge that people in general are less different from each other than we usually think. It makes the problem of picking likely winners – or, simply put, people with the winning potential – incredibly difficult. It is a challenge that businesses have attempted to solve for many years, and the best that the scientific literature on personnel management has come up with so far is the trite statement that past performance is the best predictor of future performance. Although it is true, are there any other factors that might hypothetically help to identify best performers? Thus, the goal of the project is to distinguish such factors and describe a profile of a performer that has the winning potential based on the data from CrossFit Open Games 2019/2020. By combining performance metrics with demographic data and other features, we will focus on picking not the winners (who are few), but people who might develop into winners, given the right level of investment and leadership (who are many).

In the past, CrossFit competitions have been examined by data scientists for different purposes, but we haven't seen yet high-quality predictions of winners so far. With this amount of data, CrossFit is an interesting playground to exercise data science techniques.

We used the CrossFit data because it provides information points that are comparable across years; there are millions of observations, and CrossFit as a discipline is the type of sport that is analogous to business: it allows for enduring high performance, with the oldest competitors being in their 80s, and it doesn't favour short-termism of any kind. In this regard, CrossFit is a much closer approximation of 'business' than any other competitive sport in which people train hard but retire young because they cannot sustain the competition with developing younger top performers. This factor makes the model easily adaptable to any types of high-performance businesses, i.e., consulting, investment banking, trading etc.

Data Selection, Pre-Processing and Wrangling

For this research, we examined a collection of data sets from Kaggle containing the results of web-scraping activities from the website of CrossFit Games (https://www.kaggle.com/jeanmidev/crossfit-games). The sets include the results from the Open phase of the competition for 2019 and 2020, as well as the score information and timing. In total, the research was based on 4 data sets (two with athletes' demographic data and two with their achievements at each stage of the competition). After visually expecting the data, we concluded that it is reliable. However, it contains multiple missing values because CrossFit organisation's website is not strict in recording the demographic data which is being collected from athletes directly. Some values are either missing or presented in a wrong format, e.g., weight in lbs vs. kgs, etc. The missing values were present in the critical fields, such as the 'weight', 'height' etc. The columns containing scores and time were inconsistent in terms of formatting, which required effort to normalise.

It constituted a significant challenge to us at this stage, that we managed to overcome by applying our knowledge of the field and correcting for the obviously flawed values through normalisation techniques. From these initial datasets, we created four clean datasets that lay foundation for our analysis, data wrangling and modelling (see below).

Table 1. Data Sets for Data Wrangling and Modelling (.csv files attached to the report)

Data Set	Shape_Beginning	Shape_Final
1_2019_open_athletes	572,653 rows × 19 cols	572,653 rows × 19 cols
2_2019_open_scores	2,863,265 rows × 13 cols	2,863,265 rows × 13 cols
3_2020_open_athletes	393,535 rows × 19 cols	393,535 rows × 18 cols
4_2020_open_scores	1,967,675 rows × 13 cols	1,967,675 rows × 13 cols

The next stage of EDA included compiling the 4 datasets into 1 training and 1 test datasets (for each gender separately). That required creating a single dataset for a single year. Thus, the 2019/2020 scores and 'time' were added to the athletes' demographic data for each year. Then, all the data was presented numerically, for which purpose categorical columns had to be converted into dummy variables. Other numerical variables have been scaled. The 'postcompstatus' (accepted/not accepted) became the target variable, and the rest of the columns – became features. The models will be trained based on the data from 2019, while the data from 2020 will be using for testing.

EDA and Feature Engineering

We missed data about the athletes' past performance, which, as we know, quite well defines future performance. We took a round number of 'top 300' athletes of all genders (which is slightly less than 0.01% of the participants) from 2019 (knowing their results as of 2021) and gave them a '1' for 2019 and 2020 while the rest of the athletes were given '0' for both years. Then, we wrapped this string of numbers into a new feature – a new column – that might allow ML algorithms to learn about the athletes past performance.

Another set of features that we created is 'top_affiliate' and 'top_country' which represent the athletes' allegiance to the gyms that have a history of making top performers in the past and indicate if the athletes come from one of the most represented countries (we chose 10 most represented countries and Iceland, 0.6% of whose population have participated in the opens and that produced many winners to the games).
Since we face an 'unsurpassable' class imbalance, we used an over-sampling technique to balance the classes in the training set. The final training and test datasets contained of a combination of scaled data and non-scaled data (dummies). See Tab. 2.

Table 2. Test and Train Datasets for Men and Women

Set Type	Rows_Men	Columns_Men	Rows_Women	Columns_Women
X_test (2020)	133,874	16	94,157	16
y_test (2020)	133,874	1	94,157	1
X_train (2019) resampled	390,754	16	292,480	16
y_train (2019) resampled	390,754	1	292,480	1

Attempts to Model the Data

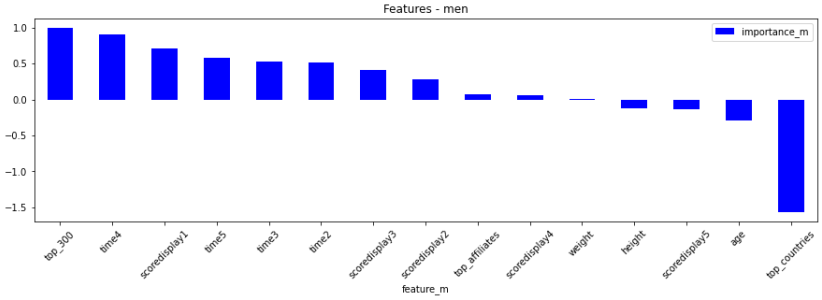
The Objective. Since human beings display little variance, we might probably not be able to achieve high accuracy in identifying the class 1 representatives – those who will get to the games. However, a high recall in this class may be the right objective to pursue because it might permit us to answer the **business question** – how to choose people with a winning potential. It is reasonable to assume that any value in the upper quintile (top 20%) might indicate a decent recall level to strive for. It is a value large enough to show that the model selected the true positives not by chance. The empirical justification for setting the 80%+ threshold for the recall parameter can be found in education: across various disciplines, 80%+ score would be indicative of either good or excellent performance and award highest grades accordingly.

Models. As evidenced by our attempts, a KNN model due to its properties won't allow us to achieve the objective. Neither will Decision Tree or Random Forest models. We've concluded that these models aren't a good fit for the modelling task at hand given the type of data that we have.

Logistic Regression: Findings

We attempted to use a logistic regression model to achieve the objective – to recall 80%+ of the class 1 representatives. Using the brute-force search, we found the regularisation parameter at 0.001 level to suit our model best for all genders. At the first attempt, we achieved an incredibly high recall score (98% for men and 94% for women) with the accuracy of ~95%. *We cannot trust the accuracy score of this model, however, because of the enormous class imbalance in the test set. This incredible recall was achieved with multiple false positives, but since we don't optimise for precision, we aren't concerned with this.*

At the next stage, we attempted to refine the model and evaluate the coefficients for statistical significance. Using the t-test, we found that all the features are statistically significant at 0.99 confidence level. In the process manifested itself a problem of multicollinearity, which we fixed by dropping the 'time1' feature for all genders. The second logistic regression demonstrated a high pseudo-R (> 80%) with the remaining 14 features.



Then, we produced another logistic regression using 'scikit-learn' and plotted the coefficients with the features they represent to analyse their relative importance. Fig. 3 below shows this importance for men. A graph for women would look similar.

In short, we found that for both men and women, the most impactful feature to define the high potential athletes are 'top_300', 'time4' (the time for w/o – here, 'workout' – 4), 'scoredisplay1' (the score for w/o 1),

'time5' (w/o 5), 'time3' (w/o 3), 'time2' (w/o 2), 'scoredisplay2' (w/o 2), 'scoredisplay3' (w/o 3), 'top_affiliates' (being a member of a gym chain that other successful athletes attended in the past), 'weight' (obviously, a proxy for muscle mass). The factors that reduce the chances to get to the games are, predictably 'age', 'height' (may affect the speed and agility), 'scoredisplay5' (w/o 5, with the score being capped, and thus achieved by many), 'top_country' (because of a high competition from compatriots).

To gain a deeper understanding of underlying causes for high/low chances to get to the games, we performed a PCA with 3 PCs explaining 80%+ of variance. We wanted to reduce the number of components to the bare minimum and simplify the picture. The surface level **recommendation** will be for the athletes to optimise their training in such a way that they increase the metrics maximising their chances to get to the games and minimise the behaviours and factors that reduce such chances (see the project Notebook).

Based on the PCA data with 3 PCs, the regression performs equally well regardless of the loss of variance: the recall for men is 99% and for women – 97%. The model started showing a different level of accuracy for both men and women, 73% and 76% respectively, which is lower than before, but because it is lower, it seems more meaningful. For men and women, 40%+ of variance is explained by PC1 and PC3, while PC2 explains around a third of variance. We've found that to improve the athletes' chances, they should maximise PC1 and PC3, which can be achieved by maximising their performance in the features that increase the values of these components, and if not, they should not fall too far away from the mean performance, to keep PC2 value low, as it minimises the athletes' chances to get to the games. See Notebook for more detail.

Likely Next Steps

The potential next steps could be to increase the accuracy of the prediction. It can be achieved by adding more features that describe the previous performance, such as scores in the past games for as many years as possible, perhaps expressed as percentile (to account for the differences in the constituent parts of the workouts across the years), their history of participation in the previous Opens and/or Games, their team performance (if known), etc. This might help to reduce the number of 'falls positive' picks.

Then, we might attempt to predict athletes' scores and relative positions in the actual Games, which can be achieved through ranking them based on the total scores predicted through a linear regression model or another method. We could investigate the potential of artificial neural networks in picking the athletes that would get through the Open stage and also in identifying their likely ranks in the games.

Business Recommendations – 4 Business Use Cases

Use Case 1: Consulting. The proposed approach (or, model) to identifying athletes with winning potential can be widely accepted in many organisations that rely on top performers. The approach allows to confidently select a large group of individuals, containing almost everyone who theoretically may become a 'star', accompanied by many other performers ('falls positives', in our case) who may still be classified as people with winning potential, because they demonstrate 'winning attributes' and can be 'converted' into stars by tuning some of those attributes.

Use Case 2: CrossFit. In the CrossFit case specifically, this approach may be used by gym owners, team leaders and coaches, who want to maximise the performance of their athletes and, cumulatively, their 'teams'. It will allow the coaches to make better decisions and not rely solely on their intuition in selecting the athletes. They will be able to tell more confidently who has got the winning potential and who hasn't.

Use Case 3: Military. The approach might be embraced by the military for the formation of super-performers, or super-performer teams. In the army, teamwork is extremely important, but it can be easily hindered by poor individual performance, should a wrong choice be made. This approach might be helpful in not only judging the level of individual physical performance, but also their character (for which the proxies might be their previous ranks in the top levels), and this approach is good for selecting large groups of people, which is specifically useful for mass drafts.

Use Case 4: HR. This approach demonstrates that there is little sufficient material to justify or explain why some people are believed to have talents that are superior to those of others. As we know, there is generally little variance between people and within their features, which indicates that almost everyone can do almost everything. That invalidates most of the modern approaches to selecting personnel and may help to build corporate or government policies regarding HR, training etc., which restrict the attempts of some organisations and departments within organisations to engage in endless greedy search for 'the best people'. The 'best people' are only very broadly and vaguely distinguishable from the rest of the pack, and at the top levels of performance, it is hardly possible to identify the best of the best confidently and accurately, hence why should one try? This can be used as an argument that the best people can be 'made', and they don't have to be greedily looked for, because this 'search' may be endless, fruitless, and wrong choices can be made based on inaccurate and misleading interpretations.